Seminar: Efficient and Robust Natural Language Processing

Seminar Overview

Large Language Models (LLMs) achieve impressive results across a wide variety of NLP tasks and languages. This increase in performance comes at a cost: better models typically require more parameters, more training data, more memory, more energy, and longer inference times, making both training and deployment prohibitively expensive at scale.

This seminar asks: how can we build NLP systems that are not only powerful, but efficient and robust? We explore this question across five thematic parts:

- **Efficient architecture:** how model design choices such as Mixture-of-Experts routing and weight pruning reduce the computational cost of large models from the ground up.
- **Efficient fine-tuning (PEFT):** how to adapt large pre-trained models to new tasks without updating all parameters, covering adapters, prompt tuning, LoRA variants, and memory-efficient training strategies.
- **Data efficiency:** how to select, curate, and generate training data more intelligently, including pre-training data curation at scale, active learning for fine-tuning, and synthetic data generation via self-play and distillation.
- **Efficient inference:** how to reduce the cost of running models at deployment time through IO-aware attention algorithms, speculative decoding, multi-head decoding, and prompt compression.
- **Robustness:** how robust are efficient LLMs to adversarial manipulation and distribution shift, and what trade-offs arise between efficiency and robustness?

The seminar consists of ten thematic sessions across these five parts. Students present a paper of their choice from each session's paper pool and lead the class discussion. The paper pool spans foundational work and the latest results from ACL, EMNLP, NeurIPS, and ICML venues.

How to Read the Schedule

Each session has a pool of 3-4 papers, there will be two presentations per session. Students choose one paper from the pool as their main presentation paper. The remaining papers in the pool are expected background, i.e. you should be able to speak to them during discussion even if you did not present them.

Papers marked [background reading] are required reading before the session but are not assigned as presentation papers. They provide the vocabulary needed to understand the presented work.

Session Schedule

#	Part	Topic	Session Description	Paper Pool
Part 0 — Foundations				
01	Foundations	**Introduction: The Efficiency & Robustness Landscape**	Instructor session. We introduce the seminar, establish a shared vocabulary, and frame the problem: why are LLMs expensive, and what does it mean to make them more efficient or more robust? Topics include the anatomy of LLM cost (parameters, memory, data, inference latency), an overview of scaling laws and their limits, a taxonomy of the efficiency approaches covered in the	Instructor-led session — no paper presentation no preparation required beyond the BERT and Transformer prerequisites

#	Part	Topic	Session Description	Paper Pool
			seminar (architecture, fine-tuning, data, inference), and a first look at the robustness challenge and why it is inseparable from the efficiency agenda.	
Part 1 — Efficient Architecture				
02	Efficient Architecture	**MoE, Sparsity & Compression**	Architecture choices set the cost ceiling for everything downstream. This session covers routing sparsity via Mixture-of-Experts, weight sparsity through one-shot pruning, and post-training quantization. As a contrast, two papers from the pool examine what models can achieve without large-scale pretraining at all, establishing a lower-bound baseline that motivates the efficiency agenda.	**#1** DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models (ACL 2024) → **#2** SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot (ICML 2023) → **#3** OneBit: Towards Extremely Low-bit Large Language Models → **#4** LRQuant: Learnable and Robust Post-Training Quantization for Large Language Models (ACL 2024) → **#5** Exploring Transformers as Compact, Data-efficient Language Models (CoNLL 2023) → **#6** NLP From Scratch Without Large-Scale Pretraining (ICML 2022) →
Part 2 — Efficient Fine-Tuning (PEFT)				
03	PEFT I	**Adapters, Prompt Tuning & Multi-task PEFT**	The first of two sessions on parameter-efficient fine-tuning (PEFT). We cover the two main paradigms (adapter-based methods and prompt/prefix tuning) and see how they extend to multi-task settings. The Houlsby et al. adapter paper is assigned as required background reading before the session.	**#7** Parameter-Efficient Transfer Learning for NLP (Houlsby et al., ICML 2019) [background reading] → **#8** The Power of Scale for Parameter-Efficient Prompt Tuning (EMNLP 2021) → **#9** ATTEMPT: Parameter-Efficient Multi-task Tuning via Attentional Mixtures of Soft Prompts (EMNLP 2022) → **#10** Hyperdecoders: Instance-specific Decoders for Multi-task NLP (EMNLP 2022) →
04	PEFT II	**LoRA Variants, Search & Memory-Efficient Training**	The second PEFT session asks: how do we choose, prune, or improve LoRA configurations — and can we do better than LoRA entirely?	**#11** GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection (ICML 2024 Oral) → **#12** AutoPEFT: Automatic Configuration Search for Parameter-

#	Part	Topic	Session Description	Paper Pool
				Efficient Fine-Tuning (TACL 2024) → **#13** LoRA-drop: Efficient LoRA Parameter Pruning based on Output Evaluation (COLING 2025) → **#14** Quantized Side Tuning: Fast and Memory-Efficient Tuning of Quantized Large Language Models (ACL 2024) →
Part 3 — Data Efficiency				
05	Data Efficiency I	**Data Curation, Selection & Instruction Tuning**	Data quality is as important as model design. This session examines efficiency from the data side at both pre-training scale (which web data do you keep?) and fine-tuning scale (which labelled examples matter most?), including active learning approaches that minimise annotation effort.	**#15** DataComp-LM (DCLM): In Search of the Next Generation of Training Sets for Language Models (NeurIPS 2024) → **#16** How to Train Data-Efficient LLMs → **#17** Rephrasing the Web: A Recipe for Compute and Data-Efficient Language Modeling (ACL 2024) → **#18** FreeAL: Towards Human-Free Active Learning in the Era of Large Language Models (EMNLP 2023) →
06	Data Efficiency II	**Synthetic Data, Self-Improvement & Distillation**	If labelled data is scarce, generate your own. This session covers three strategies: self-play (a model generates its own training signal by competing against earlier versions), knowledge distillation with rationales (a smaller model learns step-by-step reasoning from a larger one), and externally guided synthetic instruction generation.	**#19** Self-Play Fine-Tuning (SPIN): Converts Weak Language Models to Strong Language Models (ICML 2024) → **#20** Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes (ACL 2023) → **#21** RECOST: External Knowledge Guided Data-efficient Instruction Tuning (ACL Findings 2024) →
Part 4 — Efficient Inference				
07	Efficient Inference	**Decoding, Quantization & Attention Efficiency**	Training efficiency is only half the picture: deployment cost matters too. This session covers the full inference stack: IO-aware exact attention (FlashAttention-2), accelerated token generation via speculative decoding and multi-head decoding, and input-side compression via learnable prompt summarisation.	**#22** FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning (ICLR 2024) → **#23** MEDUSA: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads (ICML 2024) →

#	Part	Topic	Session Description	Paper Pool
				#24 Learning to Compress Prompts with Gist Tokens (NeurIPS 2023) →
Part 5 — Robustness & Synthesis				
08	Robustness I	**Adversarial Attacks & Instruction-Following Robustness**	How robust are LLMs to adversarial manipulation? This session introduces adversarial attacks specifically targeting LLMs in deployment: attacks on zero-shot evaluation and judging pipelines, failures in instruction following under adversarial perturbation, and robustness of vision-language models. A key discussion question: do efficient models (quantised, pruned, PEFT-tuned) degrade faster under pressure than their full counterparts?	**#25** Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment (EMNLP 2024) → **#26** Evaluating the Instruction-Following Robustness of Large Language Models (EMNLP 2024) → **#27** Improving Adversarial Robustness in Vision-Language Models with Architecture and Prompt Design (EMNLP Findings 2024) →
09	Robustness II	**Efficiency–Robustness Trade-offs**	This session revisits papers from the efficiency arc (Parts 1–4) and asks: what does each approach sacrifice in terms of robustness, generalization, or reliability? Presenters are expected to frame their chosen paper through this critical lens, even if the paper itself does not address robustness directly.	**#28** STAR: Constraint LoRA with Dynamic Active Learning for Data-Efficient Fine-Tuning of Large Language Models (ACL Findings 2024) → **#29** InsCL: A Data-efficient Continual Learning Paradigm for Finetuning Large Language Models with Instructions (NAACL 2024) → **#30** A Study of Parameter-Efficient Fine-tuning by Learning to Efficiently Fine-Tune (EMNLP Findings 2024) →
10	Synthesis	**Synthesis/Buffer Week**		

Full Paper List

All papers are listed below in order of appearance in the syllabus. Papers without a session assignment are background reading only.

#	Session	Title	Link
Part 1 — Efficient Architecture			
1	S02	DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models (ACL 2024)	aclanthology.org
2	S02	SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot	arxiv.org
3	S02	OneBit: Towards Extremely Low-bit Large Language Models	arxiv.org

#	Session	Title	Link
4	S02	LRQuant: Learnable and Robust Post-Training Quantization for Large Language Models (ACL 2024)	aclanthology.org
5	S02	Exploring Transformers as Compact, Data-efficient Language Models (CoNLL 2023) — alternative: efficiency without scale	aclanthology.org
6	S02	NLP From Scratch Without Large-Scale Pretraining (ICML 2022) — alternative: efficiency without scale	proceedings.mlr.press
Part 2 — Efficient Fine-Tuning (PEFT)			
7	S03	Parameter-Efficient Transfer Learning for NLP (Houlsby et al., ICML 2019) [background reading]	proceedings.mlr.press
8	S03	The Power of Scale for Parameter-Efficient Prompt Tuning (EMNLP 2021)	aclanthology.org
9	S03	ATTEMPT: Parameter-Efficient Multi-task Tuning via Attentional Mixtures of Soft Prompts (EMNLP 2022)	aclanthology.org
10	S03	Hyperdecoders: Instance-specific Decoders for Multi-task NLP	arxiv.org
11	S04	GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection (ICML 2024 Oral)	arxiv.org
12	S04	AutoPEFT: Automatic Configuration Search for Parameter-Efficient Fine-Tuning (TACL 2024)	aclanthology.org
13	S04	LoRA-drop: Efficient LoRA Parameter Pruning based on Output Evaluation (COLING 2025)	aclanthology.org
14	S04	Quantized Side Tuning: Fast and Memory-Efficient Tuning of Quantized Large Language Models (ACL 2024)	aclanthology.org
Part 3 — Data Efficiency			
15	S05	DataComp-LM (DCLM): In Search of the Next Generation of Training Sets for Language Models (NeurIPS 2024)	arxiv.org
16	S05	How to Train Data-Efficient LLMs	arxiv.org
17	S05	Rephrasing the Web: A Recipe for Compute and Data-Efficient Language Modeling (ACL 2024)	aclanthology.org
18	S05	FreeAL: Towards Human-Free Active Learning in the Era of Large Language Models (EMNLP 2023)	aclanthology.org
19	S06	Self-Play Fine-Tuning (SPIN): Converts Weak Language Models to Strong Language Models	arxiv.org
20	S06	Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes	arxiv.org
21	S06	RECOST: External Knowledge Guided Data-efficient Instruction Tuning (ACL Findings 2024)	aclanthology.org
Part 4 — Efficient Inference			
22	S07	FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning (ICLR 2024)	arxiv.org
23	S07	MEDUSA: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads	arxiv.org
24	S07	Learning to Compress Prompts with Gist Tokens	arxiv.org
Part 5 — Robustness & Synthesis			
25	S08	Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment (EMNLP 2024)	aclanthology.org
26	S08	Evaluating the Instruction-Following Robustness of Large Language Models (EMNLP 2024)	aclanthology.org
27	S08	Improving Adversarial Robustness in Vision-Language Models with Architecture and Prompt Design (EMNLP Findings 2024)	aclanthology.org
28	S09	STAR: Constraint LoRA with Dynamic Active Learning for Data-Efficient Fine-Tuning of Large Language Models (ACL Findings 2024)	aclanthology.org

#	Session	Title	Link
29	S09	InsCL: A Data-efficient Continual Learning Paradigm for Finetuning Large Language Models with Instructions (NAACL 2024)	aclanthology.org
30	S09	A Study of Parameter-Efficient Fine-tuning by Learning to Efficiently Fine-Tune (EMNLP Findings 2024)	aclanthology.org

Recommended Background Reading

The following papers are not assigned as session readings but provide useful context for the seminar as a whole:

- **Hoffmann et al. (2022):** Training Compute-Optimal Large Language Models (Chinchilla) — The foundational paper on compute-optimal scaling laws; essential background for understanding why data efficiency matters.
- **Strubell et al. (2019):** Energy and Policy Considerations for Deep Learning in NLP (ACL 2019) — Motivates the efficiency agenda concretely through energy consumption analysis.